Optical Engineering 49(9), 097003 (September 2010)

Discovering objects and their location in videos using spatial-temporal context words

Hao Sun
National University of Defense Technology
School of Electrical Science and Engineering
47 Yanwachi Street
Changsha, Hunan 410073, China
E-mail: clhaosun@gmail.com

Cheng Wang
Boliang Wang
Xiamen University
School of Information Science and Technology
Department of Computer Science
Xiamen, Fujian 361005, China

Naser El-Sheimy
University of Calgary
Department of Geomatics Engineering
2500 University Drive Northwest
Calgary, Alberta T2N 1N4, Canada

Abstract. We present a novel unsupervised learning algorithm for discovering objects and their location in videos from moving cameras. The videos can switch between different shots, and contain cluttered background, occlusion, camera motion, and multiple independently moving objects. We exploit both appearance consistency and spatial configuration consistency of local patches across frames for object recognition and localization. The contributions of this paper are twofold. First, we propose a combined approach for simultaneous spatial context and temporal context generation. Local video patches are extracted and described using the generated spatial-temporal context words. Second, a dynamic topic model, based on the representation of a bag of spatial-temporal context words, is introduced to learn object category models in video sequences. The proposed model can categorize and localize multiple objects in a single video. Objects leaving or entering the scene at multiple times can also be handled efficiently in the dynamic framework. Experimental results on the CamVid data set and the VISAT™ data set demonstrate the effectiveness and robustness of the proposed method.
[DOI: 10.1117/1.3488041]

Subject terms: object discovery; spatial-temporal context words; unsupervised learning; dynamic topic model.

Paper 100298R received Apr. 7, 2010; revised manuscript received Jul. 2, 2010; accepted for publication Jul. 23, 2010; published online Sep. 9, 2010.

1 Introduction

Category-level object recognition in images and videos poses a long-standing key challenge for computer vision. Over the last decade, much progress has been realized for images in scenes of limited complexity. However, the much more general and less constrainted setting of category-level object recognition in videos from cameras on a moving platform without heavy supervision during training still poses a highly ambitious computer vision task, and the required algorithms are situated at the forefront of modern vision research.[1] In this paper, we aim to recognize object categories and localize them in videos observed by moving cameras, where only unlabeled data are provided. We advocate the use of an unsupervised learning setting because it opens the possibility to take advantage of the increasing amount of available video data, without the expense of detailed human annotation. Because no labeled videos are needed for training the system and no examples are used for specifying particular objects, the task is also referred to as video object discovery.[2,3]

Video object discovery is highly interesting for a variety of applications: detecting relevant activities in surveillance video, summarizing and indexing video sequences, organizing a digital video library according to relevant objects, automatic video analysis, etc. It remains, however, a challenging problem due to cluttered background, camera motion, occlusion, viewpoint changes, and geometric and photometric variances of objects. An additional challenge is that the unknown objects can leave or enter the scene at multiple times.

In this work, we propose a generative graphical model approach to categorize and localize objects in videos, taking advantage of the robust representation of sparse spatial-temporal context words and an unsupervised learning approach. Our model involves two processes: (1) At the feature level, extracting salient video patches that are robust to pose, scale, and lighting variations, and are generic enough for dealing with different types of objects. (2) At the object level, constructing appearance and spatial configuration models of large entities by exploiting their consistency across multiple frames.

The outline of the paper is as follows. We review related work in Sec. 2. In Sec. 3 we describe our approach in more detail, including joint generation of spatial context and temporal context, video representation from spatial-temporal context vocabulary, and object discovery by a dynamic topic model. Experimental results on real-world video sequences from moving cameras are reported in Sec. 4. Finally, we conclude the paper in Sec. 5.

2 Related Work

The basic setting for video object discovery is that of category-level object recognition for multiple categories, as opposed to single-class approaches such as pedestrian detectors[4] or exemplar detection.[5,6] Categorization aims at finding all the diverse instances of a category, whereas exemplar recognition detects different views of the same ob-

0091-3286/2010/$25.00 © 2010 SPIE

ject instance. Due to the large intra-class variations, categorization is generally considered to be a harder task than exemplar detection, since a single object model has to comprise very diverse instances.

A considerable amount of previous work has addressed the problem of object categorization and localization. Early computer vision methods for object categorization attempted to build robustness to background clutter by using image segmentation as preprocessing. This naive strategy floundered on the challenges presented by bottom-up image segmentation. An efficient alternative is provided by object detection methods,[7,8] where the characteristics of objects are learned from labeled data. However, the dependence on annotated training data inevitably limits the scalability of these methods. Since objects in a video sequence can be of any type, it is also difficult to train a comprehensive object detector that covers all types.

A range of different unsupervised learning methods[2] have been proposed in the literature, including random assignment, *k*-means clustering and principal component analysis, various latent-variable models, and spectral clustering schemes. Especially, probabilistic models seem well suited for tackling the unsupervised object discovery task. Sivic et al.[9] have proposed a method that builds on probabilistic latent semantic analysis (pLSA),[10] to separate images of four distinct object categories. They later extended their work, using multiple image segments as documents, so as to better localize the objects in images.[11] Liu and Chen[12] extend the pLSA model with the integration of a temporal model so as to discover objects in video. But their method is designed for videos where only a single object will be extracted.

Without having any prior knowledge about object classes or locations, another line of work is to identify objects that occur over a period of time. Some methods observe the same scene over a long time and build a color distribution model for each pixel. Unusual objects can then be identified if some pixels observe substantial deviation from their long-term color distribution models.[13] These background modeling approaches are suitable for video surveillance with a static camera, but fail in the case of a moving camera. Some methods exploit the consistency of optical flow or spatial configuration of feature points over a period of time to discover objects.[14] However, a discriminative classifier is still needed to acquire the category labels of the discovered objects.

3 Our Approach

Given a collection of unlabeled videos, our goal is to automatically learn different classes of objects present in the data and to apply the learned model to perform object categorization and localization in a new video. Our approach is illustrated in Fig. 1. A video sequence is first represented as a collection of small video patches. Video patches are extracted from the video data by simultaneous generation of both spatial and temporal context for image regions in each frame. A dynamic latent Dirichlet allocation (LDA) 15 topic model is then proposed to discover object categories and their location in videos in an unsupervised way. In the training stage, we assume that the number of object classes is known and that no objects enter or leave the scene. However, we relax this assumption at the testing stage, where our method can handle observations containing unknown objects entering or leaving the scene at multiple times.

We seek to develop a framework for effective discovery of semantic video objects. By a video object we mean a semantically meaningful spatial-temporal entity in a video. In the spatial domain, an object usually consists of patches that coexist tightly rather than being scattered around loosely. The spatial context of patches coming from an object is far more consistent than that of patches coming from background clutter.[16] In the temporal domain, patches belonging to the same part of an object often demonstrate similar motion characteristics and exhibit consistent temporal context across frames. The use of these relational constraints, imposed by both spatial context and temporal context of local patches, can provide a richer and more discriminative representation for video object discovery.

3.1 *Joint Generation of Spatial and Temporal Context*

At the feature level, the goal is to extract candidate patches from objects that will be largely unaffected by a change in camera viewpoint, object motion, camera motion, object's scale, and scene illumination, and also will be robust to some amount of partial occlusion. We investigate two representative detectors for feature extraction in video frames: (1) the Laplacian-of-Gaussian (LoG)[5] detector, which finds peak Laplacian responses across both spatial and scale dimensions in a Gaussian scale-space image representation; (2) the maximally stable extremal region (MSER)[17] detector, which finds as image areas that are stable with respect to the change of intensity thresholds. Two different types of regions are extracted in each frame, one based on LoG interest-point neighborhoods, and one based on MSERs. It is beneficial to have more than one type of region detector because in some imaged locations a particular type of feature may not occur at all. We have the benefit of region detectors firing both at points where there is signal variation in more than one direction, and in high-contrast extended regions. The two detectors are complementary, i.e., they extract regions with different properties, and the overlap of these regions is small. Regions based on LoG interest-point neighborhoods are represented as circles, and MSER regions are represented by ellipses. An example is shown in Fig. 2(b). For a 480×360-pixel video frame the total number of regions detected is 766.

In images, there exists a strong relationship within the spatial context, which can facilitate object detection when the intrinsic local information about the object is insufficient, e.g., when the object appears on a very small scale, or when the object is interfered with by background clutter. In videos, besides the spatial context, there exists a strong relationship within the temporal context, which can help to discover salient objects. For each detected region in video frames, we define its spatial context based on region adjacency in its spatial neighborhood, and its temporal context based on similar temporal patterns across consecutive frames.

3.1.1 *Spatial context*

A neighborhood graph is constructed from the detected regions for each frame. Nodes in the graph represent detected regions. Vertices between two nodes of the graph corre-

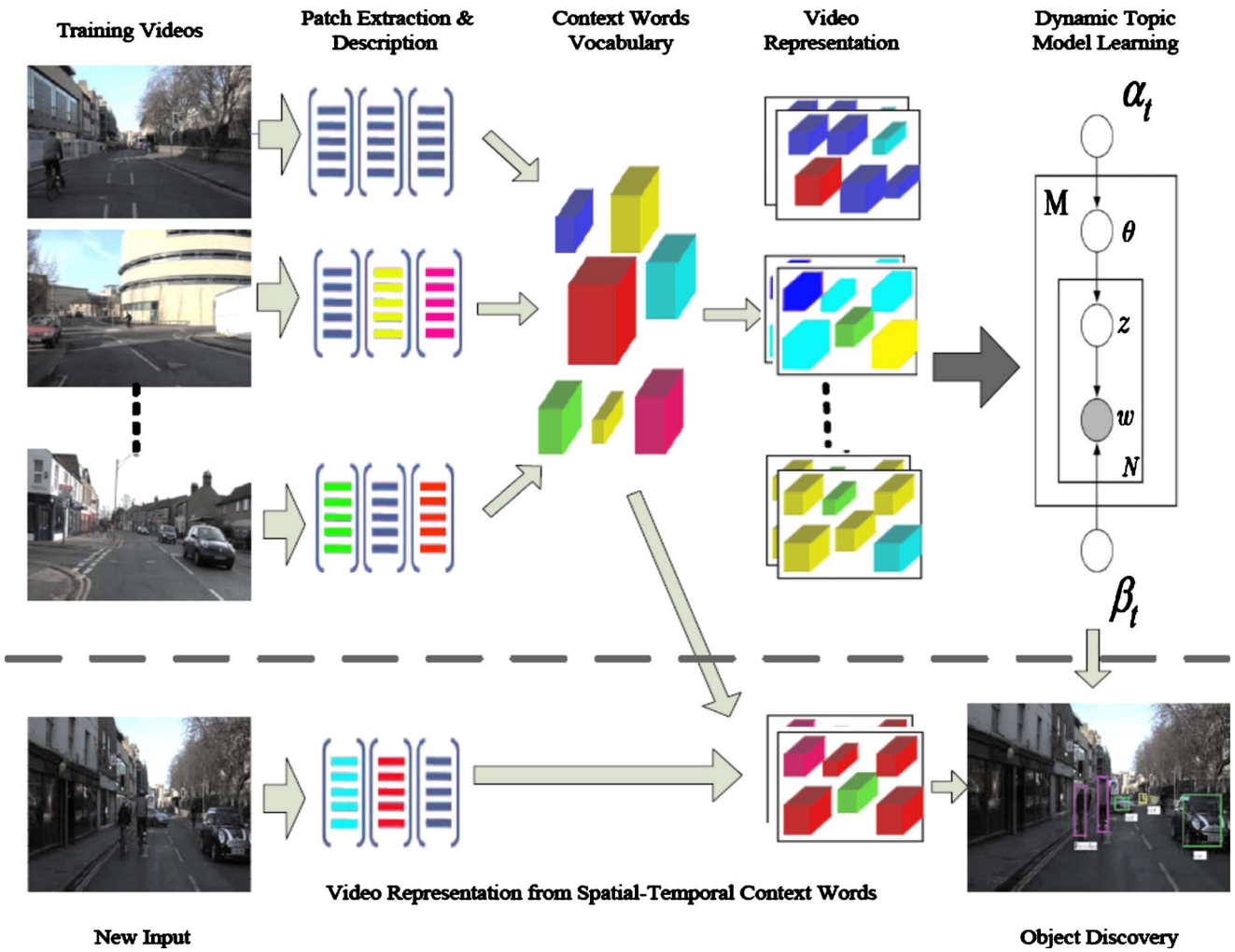


Fig. 1 Flow chart of the proposed algorithm.

spond to neighbor regions. Two types of neighborhood are often explored for spatial configuration description: the first obtained from adjacent neighbors, the second obtained from k nearest neighbors. Adjacent neighbors are the natural way to create a neighborhood; however, they are somewhat sensitive to various fluctuations. From one frame to another, the same object region may have different adjacent regions due to illumination or viewpoint changes. We adopt the k nearest neighbors (in our experiments, we use $k=3$), which are more robust to various changes, for spatial context description. Figure 2(d) shows an example of the spatial context for MSER regions detected on a frame from the CamVid data set.[18]

3.1.2 *Temporal context*

Regions detected in video frames are tracked using normalized cross correlation. Initially detected regions in the first frame are putatively matched with detected regions in the second frame, within a fixed disparity threshold of 60 pixels. Regions with similar size and scale are preferred. An intensity correlation computed over the area of a region removes all putative matches below 0.90. Motion vectors are then grouped according to their directions and magnitudes, and the acquired motion groupings are used for

Fig. 2 Spatial context for image regions: (a) a frame from the CamVid data set; (b) all detected regions superimposed on the original frame; (c) close-up of the car object with detected MSER regions superimposed; (d) the neighborhood graph of detected MSER regions on the car when $k=3$. Red color represents a graph constructed from the nearest neighbors, green color represents a graph constructed from the second-nearest neighbors, and yellow color represents a graph constructed from the third-nearest neighbors. (Color online only.)

Fig. 3 Region tracking results over several consecutive frames. The original frames are from the VISAT™ data set. Different colors of each track represent different motion vectors between adjacent frames along the trajectory.

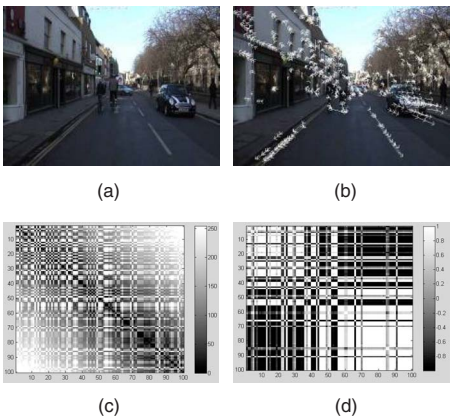


Fig. 4 Spatial-temporal context generation: (a) a frame from the CamVid data set; (b) the original frame with the LoG-based region tracks over three consecutive frames superimposed; (c) spatial adjacency matrix of selected 100 regions shown as an image; (d) temporal similarity matrix of selected 100 regions shown as an image.

tracking prediction and refinement in consecutive frames. The output is a collection of tracks $f_1,\ldots,f_n$. Each track has a start frame $s(f_i)$, an end frame $e(f_i)$, and a sequence of frame locations $f_i=\{f_i^{(t)}|s(f_i)\leqslant t\leqslant e(f_i)\}$. We require the region labels to be consistent over the entire track and denote then by l_i for track f_i, with $L=\{l_1,\ldots,l_n\}$ being the set of all track labels. Any region that does not survive for more than three frames is rejected. This stability check reduces the number of regions significantly. Figure 3 shows several LoG-based region tracking results over four or more frames, superimposed on the original frames from the VISAT™[19] data set. Different colors of each track represent different motion vectors between adjacent frames along the trajectory. For example, the red color represents motion vectors between $f_i^{(t-1)}$ and $f_i^{(t)}$ of track f_i, the green color represents motion vectors between $f_i^{(t)}$ and $f_i^{(t+1)}$ of track f_i, etc.

3.1.3 *Combining spatial context and temporal context*

Even if the two tasks of spatial context and temporal context generation are different, they are obviously related. If we have a stable spatial context for regions in each frame, it becomes easy to predict their temporal characteristics across frames. Inversely, knowing the temporal pattern of regions across frames can help to determine their spatial context. In this section, we present a strategy for joint spatial-temporal context generation, by exploiting the interdependence of appearance consistency and spatial configuration consistency of image regions across frames.

Appearance consistency of image regions across frames is enforced using temporal correspondences resulting from region tracking. Formally, this is written as a temporal similarity matrix $C_{t-1,t}$:

$$C_{t-1,t}(i,j)=\frac{p_{t-1,t}(l_i)p_{t-1,t}(l_j)}{|p_{t-1,t}(l_i)||p_{t-1,t}(l_j)|}, \quad (1)$$

where $p_{t-1,t}(l_i)$ denotes the motion vector from frame $t-1$ to frame t of the region track l_i. Spatial configuration consistency is encoded by using region adjacency in a normalized distance matrix W_{t-1}, where $W_{t-1}(i,j)$ represents the normalized block distance between the centroids of region i and region j in frame $t-1$. The indicator matrix X_t for spatial-temporal context generation can then be computed as follows:

$$X_t=W_{t-1}C_{t-1,t}, \quad (2)$$

where $X_t(i,j)$ is the probability that the region tracks l_i and l_j belong to the same object at frame t. Notice that both spatial adjacency and temporal similarity are implicitly comprised in the indicator matrix. Video patches are extracted by constructing a spatial-temporal tube that contains each LoG-based region and its top three neighbor regions from X_t, or each MSER region and its top two neighbor regions from X_t. Examples of the spatial adjacency matrix and the temporal similarity matrix shown as images are given in Fig. 4.

3.2 *Video Representation from Spatial-Temporal Context Vocabulary*

Following the preceding approach we can extract a set of meaningful video patches. Small video patches constitute the local information that is used to learn and recognize salient object categories. By employing local features, we intend to emphasize the importance and distinctiveness of the short-range spatial-temporal patterns. We argue that the observed local patterns are discriminative enough across object classes, and provide a reasonable feature space that allows building good models.

For each region within the video patch, its SIFT descriptor[5] is computed, and all the computed descriptors are then concatenated to form the appearance descriptor. This descriptor is then projected to a 64-dimensional space using a principal-component analysis (PCA) dimensionality reduction technique. Correlation ratios of region temporal patterns within the video patch are used for computing the configuration descriptor. The final descriptor for each video patch is obtained by concatenating its appearance and configuration descriptors. Normalization of the descriptors

makes them robust to contrast changes or scale changes.[2] In our experiments, we use the L^2 norm to perform a global normalization of the video descriptor.

In order to learn the vocabulary of spatial-temporal context words, we consider the set of descriptors extracted from video patches in the training data. The vocabulary is constructed by clustering using the k-means algorithm and Euclidean distance as the clustering metric. The center of each resulting cluster is defined to be a spatial-temporal context word. Thus, each video patch can be assigned a unique visual word, and a video can be represented as a collection of spatial-temporal context words from the vocabulary.

3.3 *Object Discovery by a Dynamic LDA Model*

Topic models[10,15,20] have been used in text and linguistic domains for automatically discovering topics from a collection of documents. Recently, topic models have been applied to unsupervised object discovery in images and have shown promising results.[21,22] In this work, we apply topic models to video object discovery. Object categories are treated as topics, and visual words are acquired by quantizing the spatial-temporal context descriptor of local video patches.

LDA[15] is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. LDA assumes the following generative process for each document in a corpus D:

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose topic proportions $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n:

   a. Choose a topic assignment $z_n \sim \text{Multinomial}(\theta)$.
   b. Choose a word w_n from $p(w_n|z_n,\beta)$, a multinomial probability conditioned on the topic z_n.

This process implicitly assumes that the documents are drawn exchangeably from the same set of topics. For object discovery in videos observed by moving cameras, where unknown object categories may enter or leave the scene at multiple times, we cannot directly apply LDA to our problem, because an evolving set of object categories exist in the video data.

Note that the video data can be divided by time slice. In this work, time slice refers to the time duration where a fixed number of object categories exist in the video. We model the documents of each slice with a K-component topic model, where the topics associated with slice t evolve from the topics associated with slice $t-1$. The dynamics of topics and topic proportion distributions is modeled as

$$\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1},\delta^2 I), \qquad \beta_t|\beta_{t-1} \sim N(\beta_{t-1},\delta^2 I). \tag{3}$$

Suppose we have a set of M $(j=1,\ldots,M)$ video sequences containing spatial-temporal context words from a vocabulary of size V $(i=1,\ldots,V)$. Each video d_j is divided into T_j slices and represented as a sequence of N_j spatial-temporal context words $w=(w_1,w_2,\ldots,w_{N_j})$. The generative process for slice t $(t=1,\ldots,T_j)$ is thus as follows:

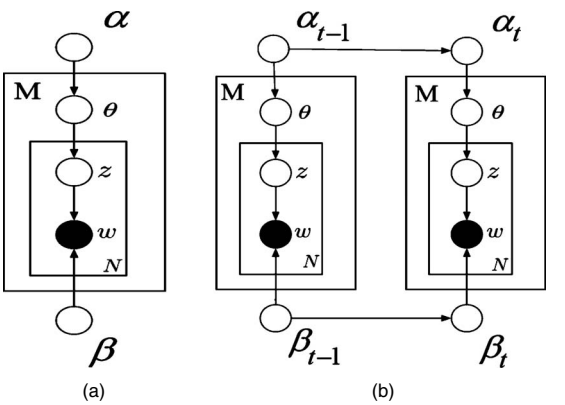


Fig. 5 Graphical models. (a) Latent Dirichlet allocation (LDA) graphical model. Nodes are random variables. Shaded ones are observed, and unshaded ones are unobserved. The plates indicate repetitions. (b) Graphical model that represents the dynamic LDA model. The dynamic LDA model models an evolving set of topics.

1. Draw topics $\beta_t|\beta_{t-1} \sim N(\beta_{t-1},\delta^2 I)$.
2. Draw $\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1},\delta^2 I)$.
3. Choose the number of spatial-temporal words: $N_j \sim \text{Poisson}(\xi)$.
4. Choose topic proportions $\theta_t \sim \text{Dir}(\alpha_t)$.
5. For each of the N_j words w_n:

   a. Choose a topic assignment $z_n \sim \text{Multinomial}(\theta_t)$.
   b. Choose a word w_n from $p(w_n|z_n,\beta_t)$, a multinomial probability conditioned on the topic z_n.

Figure 5 shows the graphical model of the LDA topic model and the proposed dynamic LDA topic model. The joint distribution of a topic mixture θ_t, the set of words w observed in the current video slice, and their corresponding topic (object category) z can then be written as

$$p(\theta_t,z,w|\alpha_t,\beta_t) = p(\theta_t|\alpha_t)\prod_{n=1}^{N} p(z_n|\theta_t)p(w_n|z_n,\beta_t). \tag{4}$$

Given a new input, the posterior distribution of the hidden variables is computed as

$$p(\theta_t,z|w,\alpha_t,\beta_t) = \frac{p(\theta_t,z,w|\alpha_t,\beta_t)}{p(w|\alpha_t,\beta_t)}, \tag{5}$$

where θ_t is specific to each input and represents its latent topic distribution. Although it is computationally intractable to perform inference and parameter estimation for the dynamic that model, several approximation algorithms for that model have been investigated. In our experiments, we use the variational inference approach proposed in Ref. 15. The family of variational distributions is characterized by

$$q(\theta,z|\gamma,\phi) = q(\theta|\gamma)\prod_{n=1}^{N} q(z_n|\phi_n), \tag{6}$$

where γ and θ are the free variational parameters. The corresponding optimization procedure produces the parameters (γ^*,ϕ^*), which are functions of w.

Fig. 6 Example images from video sequences of the CamVid data set (first column) and the VISAT™ data set (other columns).

Once θ_t is inferred, we can classify video sequences by selecting the most likely topics (object categories) in the current testing video. Furthermore, we are also interested in localizing multiple objects in a single video sequence. Each video patch can be labeled with an object category by selecting the topics that generates its corresponding spatial-temporal context word with highest probability. Thus, we label the regions and their spatial-temporal neighborhood that support the video patches, effectively producing object localization. A spatial clustering algorithm can also be applied to the support regions of particular objects. The bounding boxes and relevant characterization of every object, including the centroid position, size, and number of correspondences on it, can be extracted.

4 Experimental Results

We test our algorithm using two data sets: the CamVid data set[18] and the VISAT[19] data set. The CamVid data set depicts moving driving scenes in the city of Cambridge (UK), filmed from a moving car. The VISAT data set depicts scenes in and around the city of Calgary, filmed from multiple cameras on a mobile mapping platform. The resolutions of the CamVid and VISAT images are 960×720 and 1600×1200 pixels, respectively. In our experiments, all frames from both data sets are downsized to 480×360 pixels for process efficiency. Both data sets contain videos of cluttered background, occlusions, moving cameras, moving cars, and other independently moving objects. Figure 6 shows example images from video sequences of the data sets.

For simplicity, in our experiments we focus on the major moving-object categories in video sequences from moving cameras. The dynamic LDA model is fitted to four classes: pedestrian, bicyclist, car, and moving background. We extract local video patches with the procedure described in Sec. 3.1 and describe the corresponding spatial-temporal tubes using both appearance descriptors and spatial configuration descriptors. In order to build the spatial-temporal context vocabulary, video patch descriptors computed from the training videos are clustered. To understand how the algorithms perform, we train on video collections for which we know the desired visual topics.

The latent topic model provides means to rank the spatial-temporal context words, given an object class. Visual words that are most probable for the four discovered topics are shown in Fig. 7. Topic discovery analysis cleanly separates local video patches into different object classes. Increasing the number of topics in the moving background will help to discover more object classes, e.g., building, wall, tree, sidewalk. The most likely words for each discovered topic appear to be semantically meaningful patches.

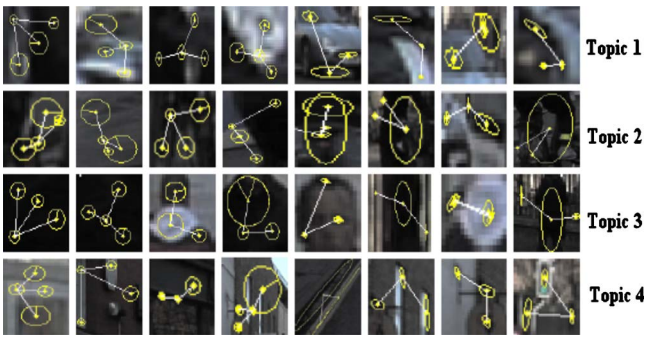


Fig. 7 The most likely spatial-temporal context words (shown by eight examples in a row, four examples of each feature type) for the four learnt topics in our experiments. The first row shows words corresponding to the car topic, the second row to the bicyclist topic, the third row to the pedestrian topic, and the fourth row to the moving background topic.

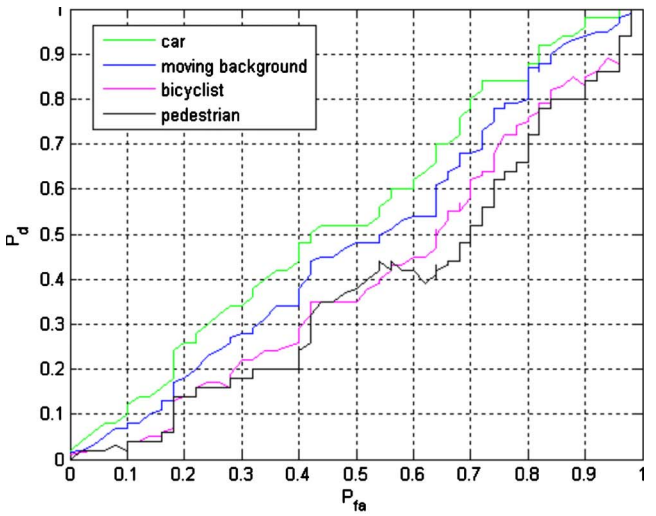
We run our experiments on a Pentium Dual-Core Machine with 2.60-GHz CPU. The average time to train the model is 22.3 s, using a vocabulary of 300 spatial-temporal context words.

To evaluate the recognition performance of the learned object model, the discovered topics are used for classifying video sequences by selecting the most likely topic existing in the sequences. We have collected 160 testing video sequences from the two data sets, 40 sequences for each discovered topic. Each testing video sequence contains only objects from one of the four classes. The receiver operating characteristic (ROC) and the recall precision curve (RPC) for the classification experiment are shown in Fig. 8(a) and 8(b), respectively. It can be seen that the best performance is obtained for the car class, because more distinctive patches are often detected on cars and the structure of the car class is well characterized by the spatial-temporal context words. Among the other three classes, the moving background obtains better performance than the bicyclist and the pedestrian. This can be partly explained by the fact that the most likely spatial-temporal context words for bicyclist and pedestrian, as shown in Fig. 7, capture certain background information. Furthermore, moving background is present in all video sequences, and it affects the performance for the bicyclist class and the pedestrian class. The pedestrian class performs poorly due to a combination of less distinctive spatial-temporal patches and sparse visual-word distribution in the visual vocabulary.

The confusion matrix for the classification experiments is given in Table 1. It shows large confusion between bicyclist and moving background, and between pedestrian and moving background. This can be partly explained by the fact that the most likely spatial-temporal context words for bicyclist and pedestrian, as shown in Fig. 7, capture certain background information.

Some of the object recognition and localization results from video sequences containing a single moving object are shown in Fig. 9. Results from video sequences containing multiple moving objects are shown in Fig. 10.

Object discovery results in video sequences containing a single moving object are given in Fig. 9. It can be shown



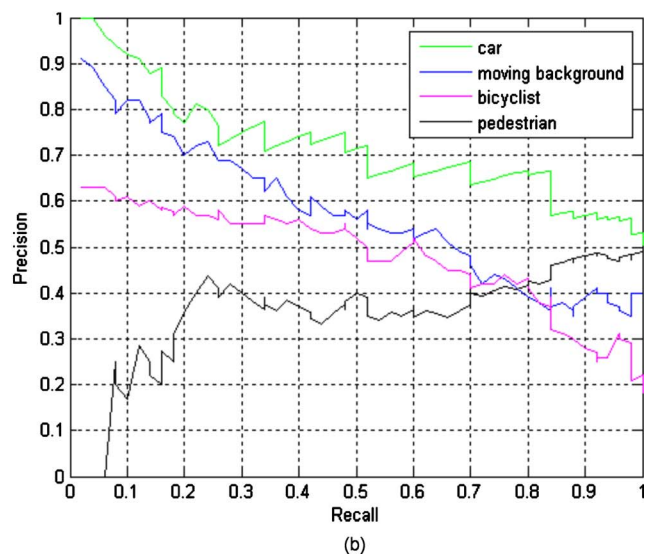


Fig. 8 The (a) ROC and (b) PRC curves for video (one topic per video) classification.

that the moving car and bicyclist are correctly recognized and localized. The first two columns show the original frames and the detected regions. Close-ups of example spatial-temporal context words on the moving object are depicted in the third column. For effective localization of

Table 1 Confusion table for video sequence classification.

True category:	Car	Bicyclist	Pedestrian	Moving background
Topic 1—car	0.87	0.03	0.02	0.08
Topic 2—bicyclist	0.05	0.74	0.04	0.17
Topic 3—pedestrian	0.07	0.08	0.72	0.13
Topic 4—moving background	0.08	0.04	0.03	0.85

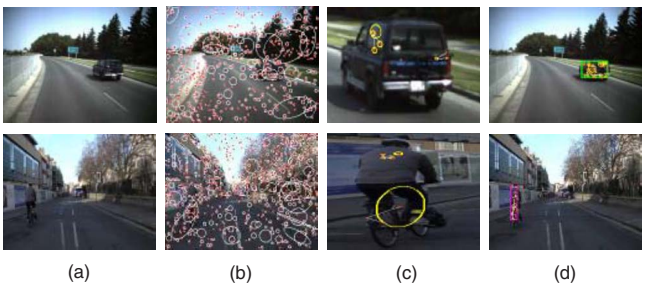


Fig. 9 Object discovery results in videos containing a single moving object: (a) the original frame from the data set; (b) frames with detected regions superimposed; (c) close-up of example spatial-temporal context words on the moving object (one for each feature type); (d) supporting regions and neighborhood labeling for object localization.

the discovered topic, spatial-temporal context words belonging to the topic with high probability and their spatial neighborhood graphs are labeled. The bounding box of each discovered object is then extracted, based on the labeled regions.

Figure 10 shows object discovery results in video sequences containing multiple moving objects. The first row shows results on a video sequence from the CamVid data set. It can be seen that both the two bicyclists and the moving car, moving in opposite directions, are correctly recognized and localized. Images in other rows are results on video sequences from the VISAT data set. Moving cars driving in the same direction are discovered in the second and the fourth row. Note that in the second row some regions belonging to the road surface have been falsely labeled as the supporting neighborhood for the two moving cars. The two moving cars driving in the traffic lane in the

Fig. 10 Object discovery results in videos containing multiple moving objects: (a) the original frame from the data set; (b) frames with detected regions superimposed; (c) spatial-temporal context words of the discovered topic and their supporting region neighborhood graphs; (d) discovered objects denoted by their bounding boxes.

third row have been recognized and localized as one car, due to a combination of minor occlusion and sparse visual words on the car in the far field.

The computation time for object discovery in a new video depends greatly the scene, because the number of features detected varies in different scenes. Our MATLAB implementation of the algorithm without optimization can process a 15-frame testing video containing a single moving object in around 6 s. For video sequences containing multiple moving objects the algorithm works a little slower, because more features are often detected in the sequences.

5 Conclusions

In this paper, a novel unsupervised learning algorithm is proposed for object discovery in videos from moving cameras. The major contributions of this paper are:

1. A combined approach for simultaneous generation of spatial and temporal context is presented for video patch extraction. Both appearance consistency and spatial configuration consistency of local patches across multiple frames are exploited to find candidate object parts.
2. A dynamic LDA model is introduced for object category recognition and localization. The proposed dynamic topic model involves an evolving set of object categories, which can handle multiple moving objects of different classes entering or leaving the scene.

Experimental results on video sequences from the CamVid data set and the VISAT data set demonstrate the effectiveness and robustness of our method for video object discovery.

In future research, we will attempt to improve the proposed algorithm in the following ways. First, contour features should be extracted for more discriminative analysis of different object categories. Second, more training videos should be collected to learn the topic-specific spatial-temporal words of the dynamic LDA model. Third, emphasis should also be placed on the computation cost of the algorithm.

Acknowledgments

The authors thank the associate editor Prof. Michael Bove and the anonymous reviewers for valuable comments, which helped improve the clarity of the presentation of this paper. The work was supported by the National Natural Science Foundation of China (Project No.40971245).

References

1. B. Ommer, T. Mader, and J. M. Buhmann, "Seeing the objects behind the dots: recognition in videos from a moving camera," *Int. J. Comput. Vis.* **83**, 57–71 (2009).
2. T. Tuytelaar, C. H. Lampert, M. B. Blaschko, and W. Buntine, "Unsupervised object discovery: a comparison," *Int. J. Comput. Vis.* **88**, 284–302 (2010).
3. D. Liu and T. Chen, "DISCOV: A framework for discovering objects in video," *IEEE Trans. Multimedia* **10**, 200–208 (2008).
4. N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. 9th Eur. Conf. on Computer Vision*, pp. 428–441 (2006).
5. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.* **60**, 91–110 (2004).
6. J. Sivic, F. Schaffalitzky, and A. Zisserman, "Object level grouping for video shots," *Int. J. Comput. Vis.* **67**, 189–210 (2006).
7. A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *Proc. IEEE Int. Conf. on Computer Vision*, pp. 606–613 (2009).
8. H. Harzallah, F. Jurie, and C. Schmid, "Combining efficient object localization and image classification," in *Proc. IEEE Int. Conf. on Computer Vision*, pp. 237–244 (2009).
9. J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering objects and their location in images," in *Proc. 10th IEEE Int. Conf. Computer Vision*, pp. 370–377 (2005).
10. T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.* **42**, 177–196 (2001).
11. B. Russell, A. Efros, J. Sivic, W. Freeman, and A. Zisserman, "Using multiple segmentations to discover objects and their extent in image collections," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1605–1614 (2006).
12. D. Liu and T. Chen, "A topic-motion model for unsupervised video object discovery," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1–8 (2007).
13. A. Mittal and N. Paragios, "Motion-based background subtraction using adaptive kernel density estimation," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 302–309 (2004).
14. M. Leordeanu and R. Collins, "Unsupervised learning of object features from video sequences," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1142–1149 (2005).
15. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
16. D. Liu and T. Chen, "Unsupervised image categorization and object localization using topic models and correspondences between images," in *Proc. 11th IEEE Int. Conf. on Computer Vision*, pp. 1–7 (2007).
17. J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proc. 13th Br. Machine Vision Conf.*, pp. 384–393 (2002).
18. G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: a high-definition ground truth database," *Pattern Recogn. Lett.* **30**, 88–97 (2008).
19. N. El-Sheimy and K. Schwarz, "Navigating urban areas by VISAT—a mobile mapping system integrating GPS/INS/digital cameras for GIS application," *Navigation* **45**, 275–286 (1999).
20. D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proc. 23rd Int. Conf. on Machine Learning*, pp. 113–120 (2006).
21. J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vis.* **79**, 299–318 (2008).
22. X. Wang, X. Ma, and W. E. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models," *IEEE Trans. Pattern Anal. Mach. Intell.* **31**, 539–555 (2009).

Hao Sun received the BSc and MSc degrees at the National University of Defense Technology, Changsha, China, in 2006 and 2008, respectively. He is currently pursuing the PhD degree in the School of Electrical Science and Engineering, National University of Defense Technology. His research interests include image analysis and understanding, pattern recognition, and information fusion.

Cheng Wang received the BSc and PhD degrees in communication and signal processing from the National University of Defense Technology, Changsha, China, in 1997 and 2002, respectively. He is currently an associate professor in the School of Electronic Science and Engineering, NUDT. He is a professor in the Department of Computer Science at Xiamen University, Fujian, China. He is also the co-chair of the Working Group I/3 "Multi-Platform Multi-Sensor Inter-Calibration" in the International Society of Remote Sensing (ISPRS). His research interests include image analysis, information fusion, and mobile mapping data processing.

Boliang Wang is a professor in the Department of Computer Science at Xiamen University, Fujian, China. He is also a professor in the School of Electronic Science and Engineering, NUDT. His area of expertise includes image processing, multisensor integration, and pattern recognition. He has published more than 80 scientific papers and patents.

Naser El-Sheimy is the head of the Department of Geomatics Engineering and leader of the Mobile Multi-sensor Research Group at the University of Calgary, Alberta, Canada. His area of expertise is in the integration of GPS/INS/imaging sensors for mapping and GIS applications with special emphasis on the use of multiple sensors in mobile mapping systems. He is the chair of the International Society for Photogrammetry and Remote Sensing Working Group (ISPRS) on Integrated Mobile Mapping Systems, the chair of the special study group for mobile multi-sensor systems of the International Association of Geodesy (IAG), and the chairman of the International Federation of Surveyors (FIG) Working Group C5.3 on integrated positioning, navigation, and mapping Systems.